

Legumain is an independent predictor for invasive recurrence in breast ductal carcinoma in situ

Michael S Toss^{1,2}, Islam M Miligy^{1,3}, Kylie L Gorringer^{4,5}, McCaffrey L⁶, Abdulbaqi AlKawaz^{1,7}, Asima Abidi¹, Ian O Ellis¹, Andrew R Green¹ and Emad A Rakha^{1,3}

¹Nottingham Breast Cancer Research Centre, Division of Cancer and Stem Cells, School of Medicine, The University of Nottingham, Nottingham City Hospital, UK.

²Histopathology department, South Egypt Cancer Institute, Assiut University, Egypt,

³Histopathology department, Faculty of Medicine, Menoufia University, Egypt, ⁴Cancer

Genomics Program, Peter MacCallum Cancer Centre, Melbourne, and ⁵The Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Australia,

⁶Department of Oncology, Rosalind and Morris Goodman Cancer Research Centre, McGill University, Montreal, Canada, and ⁷College of dentistry, Al Mustansiriya University, Iraq.

Correspondence:

Professor Emad Rakha

Department of Histopathology, Nottingham University Hospital NHS Trust,

City Hospital Campus, Hucknall Road, Nottingham, NG5 1PB, UK

Tel: (44) 0115-9691169, Fax: (44) 0115-9627768

Email: emad.rakha@nottingham.ac.uk, Emad.rakha@nuh.nhs.uk

Key words: DCIS, legumain, poor prognosis, invasive recurrence

Running title: Legumain expression in DCIS

Abstract

Legumain is a proteolytic enzyme and playing a role in the regulation of cell proliferation in invasive breast cancer. Studies evaluating its role in ductal carcinoma in situ (DCIS) are lacking. Here, we aimed to characterise legumain protein expression in DCIS and evaluate its prognostic significance. Legumain was assessed immunohistochemically in a tissue microarray of a well characterised cohort of DCIS (n=776 pure DCIS and n=239 DCIS associated with invasive breast cancer (DCIS-mixed)). Legumain immunoreactivity was scored in tumor cells and surrounding stroma and related to clinicopathological parameters and patient outcome. High legumain expression was observed in 23% of pure DCIS and was associated with features of high risk DCIS including higher nuclear grade, comedo necrosis, hormone receptor negativity, HER2 positivity and higher proliferation index. Legumain expression was higher in DCIS associated with invasive breast cancer than in pure DCIS ($p<0.0001$). In the DCIS-mixed cohort, the invasive component showed higher legumain expression than the DCIS component ($p<0.0001$). Legumain was an independent predictor of shorter local recurrence free interval for all recurrences ($p=0.0003$) and for invasive recurrences ($p=0.002$). When incorporated with other risk factors, legumain provided better patient risk stratification. High legumain expression is associated with poor prognosis in DCIS and could be a potential marker to predict DCIS progression to invasive disease.

INTRODUCTION

The key strategies for breast ductal carcinoma in situ (DCIS) management are to prevent its progression into invasive disease and to avoid disease recurrence, particularly invasive breast carcinoma which accounts for half of the recurrences. Identifying high risk DCIS that have potential to invade is an excellent approach towards patient risk assessment and stratification for individualized management (1). However, since the current clinicopathological parameters are inadequate to define DCIS risk precisely, identification of novel prognostic markers is necessary (2, 3). Furthermore, the newly described genetic signatures such as Oncotype DX DCIS for prediction of recurrence show controversial results and need further validation (4-9). Of note, genes included in Oncotype DX signature are mainly related to cellular proliferation and metabolism as subsequent indicators of invasion rather than invasive potential (6). In addition, currently available risk indices such as Van Nuys Prognostic Index and nomograms rely mainly on clinicopathological parameters and to lesser extent on markers related to the tumor cells with little consideration for the surrounding microenvironment (7, 10-12). With the emerging role of tumor microenvironment and the related proteins in the disease behaviour (13), identification of more robust genetic signatures incorporating the crosstalk between tumor epithelial cells and surrounding microenvironment might provide a better approach for DCIS risk assessment and hence better management.

Basement membrane degradation and stromal remodelling are fundamental steps in progression from DCIS into invasive disease. Although the key role of matrix metalloproteinases in stromal breakdown is undeniable, explanation of DCIS progression into invasive disease depending solely on them is insufficient. Studies that targeted blocking metalloproteinases action in order to prevent disease progression reported non-promising results (14, 15). Taken together, identification of novel markers that play a role in DCIS invasiveness might help in better understanding of the disease biology and risk stratification.

Legumain (LGMN) is a cysteine endopeptidase belonging to the asparaginyl endopeptidase family encoded by the *legumain* gene (16, 17). Legumain activates zymogen gelatinase A by cleavage of pro-gelatinase A, which is an important mediator of extracellular matrix degradation, thereby helping the tumor to invade and metastasize (18-20). It also activates other proteases that are key in regulating angiogenesis, growth and other related functions in tumors (21). Legumain is expressed at elevated levels in invasive breast cancer (16, 22), colorectal (23), prostate (24) and gastric carcinomas (17) and is related to poor prognosis (21). Moreover, legumain is differentially expressed between normal breast tissue and invasive breast cancer (16), however the role of legumain in DCIS has yet to be established. In this study, we aim to assess the pattern of legumain expression and its prognostic significance in a large well-annotated DCIS cohort.

MATERIAL AND METHODS

Study Cohort

A well characterised annotated cohort of DCIS including pure DCIS (n=776) and DCIS mixed with invasive breast cancer (DCIS-Mixed) (n=239) diagnosed between 1990 to 2012 at Nottingham City Hospital, Nottingham, United Kingdom was used as previously described (25). Patients' demographic data, histopathological parameters, management including post-operative radiotherapy and development of local recurrence were collected (Supplementary Table 1). Local recurrence free interval was defined as the time (in months) between 6 months after the first DCIS surgery and occurrence of ipsilateral local recurrence (either as DCIS or invasive breast cancer). Cases undergoing re-operation within the first 6 months due to close surgical margins or presence of residual disease were not counted as recurrence. Patients who developed contralateral disease following DCIS diagnosis were censored at the time of development of the contralateral cancer. Within a median follow up period of 103 months (range 6-240), 83 cases (11%) developed a recurrence in the pure DCIS cohort comprising 30 DCIS (36%) and 53 invasive cancer with or without DCIS (64%). Six recurrence events were developed after

mastectomy and 11 events after management with breast conserving surgery followed by adjuvant radiotherapy while the majority of the recurrences (n=66) occurred after breast conserving surgery alone.

Additionally, data on different molecular classes and tumor infiltrating lymphocytes density were available for the cohort (25, 26). To avoid selection bias, the DCIS-mixed cohort was selected with clinicopathological features comparable to the pure cohort regarding age at diagnosis, DCIS nuclear grade, and the presence of comedo necrosis.

Immunohistochemistry

Tissue microarrays were prepared from both cohorts. The TMA was constructed using a TMA GRAND MASTER 2.4-UG-EN MACHINE, using 1 mm punch sets. Cases with heterogeneous DCIS morphological patterns or grade were sampled from all representative areas. In addition, whole tissue sections from 20 cases comprising 10 pure DCIS and 10 DCIS-mixed cases were assessed to evaluate the pattern of legumain expression in malignant breast tissue and adjacent stroma and normal tissue.

Primary antibody specificity for rabbit polyclonal legumain antibody [ab125286, Abcam, UK] was validated using Western Blot on whole cell lysates of MCF7 and SKBR3 human breast cancer cell lines (obtained from the American Type Culture Collection; Rockville, MD, USA) as previously described (27-29). Legumain antibody was used at a dilution of 1:500, which showed a single specific band at the predicted size of 56 KDa.

Expression of legumain protein in DCIS was assessed by immunohistochemistry using the Novocastra Novolink TM Polymer Detection Systems kit (Code: RE7280-K, Leica, Biosystems, UK). Tissue microarray and full-face sections (4 µm) were stained with rabbit polyclonal legumain (dilution 1:150), incubated for 24 hours. Normal kidney tissue was used as a positive control while a negative control was carried out by omitting the primary antibody.

Scoring of legumain expression

Percentage of cells showing cytoplasmic granular/vesicular staining (16) was estimated in tumor epithelial cells and the surrounding stromal fibroblasts, separately. Cores containing <15% either tumor epithelial cells and/or stroma were excluded from the scoring. All scored cores showed representative areas of specialised stroma (within two high power fields) (30) surrounding the malignant ducts. In addition, the few cores included malignant epithelial cells only were excluded as it was difficult to differentiate between in situ or invasive process and the origin of these tumour cells. This method aimed to improve the reliability of the study and the cases excluded were random. Cases with multiple cores were scored and the average score was used for the analysis. For mixed cohort, each component, DCIS and invasive, was scored separately for the tumor epithelial cells and surrounding stroma. The cases were scored by two pathologists (MST and IMM) using a multiheaded microscope, considering the percentage of positive staining of any intensity. For dichotomization of protein expression, cut-off points for either malignant epithelial cells or stromal expression of legumain were defined according to the conducted results from X-tile bioinformatics software (Yale University, version 3.6.1) (31) based on local recurrence free interval in the pure DCIS cohort. High legumain expression within tumor epithelial cells was considered when more than 65% of tumor cells showed staining, while expression in more than 10% of the surrounding fibroblasts was considered high expression.

Analysis of *legumain* mRNA expression in breast cancer:

To emphasise the prognostic role of legumain in breast cancer and given the lack of data on the transcriptomic profiles of DCIS, *legumain* normalised mRNA expression was evaluated as a potential prognostic marker in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort dataset (32), which comprises a large (n=1980) cohort of invasive breast cancer with comprehensive molecular characterisation. Moreover, to validate the prognostic significant of legumain in breast cancer, analysis using the Breast Cancer Gene-Expression Miner v4.1 (bc-GenExMiner v4.1) database was carried out.

Statistical analysis

Statistical analyses were performed using SPSS v21 (Chicago, IL, USA) for Windows. Student's t test and analysis of variance (ANOVA) were used to correlate between *legumain* mRNA level as a continuous variable and other clinicopathological parameters in METABRIC data. Association with *legumain* mRNA expression and breast cancer specific survival was performed after dichotomisation of expression into high and low groups based on the median value.

Spearman's Rho test was used to correlate between legumain expression with the tumor epithelial and stromal cells. Association between legumain expression and clinicopathological parameters in pure DCIS was performed using Chi-square, Mann Whitney and Kruskal-Wallis tests. Wilcoxon signed rank test was used to compare the expression of legumain between DCIS component and invasive component within the DCIS-mixed cases. Univariate survival analysis against local recurrence free interval was carried out using log rank test and Kaplan Meier curves. Cox regression model was used for multivariate analysis of legumain expression for all recurrences (either DCIS or invasive breast cancer) and invasive recurrences. For all tests, a two-tailed *p*-value of less than 0.05 was considered as statistically significant.

This work obtained ethics approval by the North West – Greater Manchester Central Research Ethics Committee under the title; Nottingham Health Science Biobank (NHSB), reference number 15/NW/0685.

RESULTS

Pattern of legumain expression

The evaluation of full-face tissue sections demonstrated representative distribution of legumain expression either the tumor epithelial cells or the surrounding specialised stroma throughout the whole section, indicating representability of tissue microarrays to assess legumain expression in our cohort. Adjacent normal breast terminal duct-lobular units showed negative or very faint cytoplasmic staining of legumain. Occasional inflammatory and stromal cells were also stained in few cores. When present, legumain was expressed in the cytoplasm of the epithelial tumor cells and surrounding fibroblasts (Figure 1).

After unbiased exclusion of uninformative cores (lost cores, folded tissue during processing and staining of cores containing <15% tumor cells and/or stroma), the final number of cases suitable for scoring was 464 pure DCIS and 191 DCIS-mixed. Legumain expression showed a unimodal distribution. The median percentage of positive tumor epithelial cells was 25% in pure DCIS, 30% in the DCIS component of mixed cases, and 60% in invasive component of the latter (all showed a range between 0-100%). For stromal expression, the median percentage of positive stromal cells was 5% in pure DCIS (range 0-80%), 70% in the DCIS component of mixed cases (range 0-80%) and 90% in the invasive component of the latter (range 0-90%). Within the pure DCIS cohort, high legumain expression was observed in 23% and 44% in tumor epithelial and surrounding stromal cells; respectively. There was a positive linear correlation between expression of legumain within the epithelial cells and surrounding fibroblasts ($r=0.408$, $p<0.0001$, Spearman's correlation).

The proportion of cases with high legumain was greater in DCIS-mixed than pure DCIS, both within the tumor epithelial cells (23% of pure DCIS cases vs. 36% of DCIS mixed with invasive breast cancer, $\chi^2=11.7$, $p=0.001$) and stromal cells (44% for pure DCIS vs. 86% of DCIS mixed with invasion, $\chi^2=95.5$, $p<0.0001$). Similar results were observed when the data was analysed using a continuous scale ($p=0.049$ and $p<0.0001$, for tumor epithelial cells and stromal cells, respectively). Moreover, there was a statistically

significant difference between legumain expression within the tumor epithelial cells of the DCIS component and invasive component of DCIS-mixed cases ($p<0.0001$). Similarly, legumain staining was more frequent in the stromal fibroblasts surrounding the invasive component than those surrounding the DCIS component ($p<0.0001$) (Figure 2).

Significance of legumain expression in pure DCIS

High expression of legumain within the malignant epithelial cells and/or surrounding stromal fibroblasts in the pure DCIS was associated with various clinicopathological parameters characteristic of poor prognosis, including high nuclear grade, presence of comedo necrosis, hormonal receptor negativity, HER2 positivity, high proliferative index, and dense tumor infiltrating lymphocytes (Table 1). Analysis of continuous data of legumain expression scores showed similar results (Supplementary Table 2).

To validate the prognostic value of legumain in invasive breast cancer, the METABRIC cohort (32) was used to assess the levels of *legumain* mRNA and correlate its expression with the clinicopathological variables and outcome. Higher *legumain* mRNA level was associated with high tumor grade ($p=0.03$), lymph node metastasis ($p=0.04$), estrogen receptor negativity ($p=0.001$), HER2 positivity ($p=0.006$) in addition to shorter breast cancer specific survival ($HR=1.3$, 95% $CI=1.1-1.5$, $p=0.007$) (Supplementary Tables 3, and Supplementary Figure 1). Analysis using the Breast Cancer Gene-Expression Miner v4.1 (bc-GenExMiner v4.1) database showed that high *legumain* mRNA was associated with higher metastatic relapse and/or death ($HR=1.3$, 95% $CI=1.1-1.5$, $p=0.0001$).

Outcome analysis in pure DCIS cohort

High legumain expression within tumor epithelial cells was associated with shorter local recurrence free interval (all recurrences either as *in situ* or invasive disease) in pure DCIS ($HR=2.7$, 95% $CI=1.6-4.8$; $p=0.0002$, Figure 3). Association with shorter local recurrence free interval was observed in patients treated with breast conserving surgery without adjuvant radiotherapy ($HR=2.6$, 95% $CI=1.4-4.4$; $p=0.002$, Figure 3) however; the significant association with poor outcome was not maintained in patients treated with either mastectomy ($HR=0.9$, 95% $CI=0.1-8.8$; $p=0.9$) or breast conserving surgery followed by adjuvant radiotherapy ($HR=2.4$, 95% $CI=0.8-10.4$; $p=0.08$). Interestingly,

there was an association between high legumain expression and ipsilateral local recurrence as invasive disease ($HR=3.1$, 95% $CI=1.5-6.5$; $p=0.001$, Figure 3) particularly in patients treated with breast conserving surgery without post-operative adjuvant radiotherapy ($HR=3.3$, 95% $CI=1.5-7.3$; $p=0.004$). Supplementary Figure 2 shows forest plots illustrating the hazard ratio for disease recurrence of the different clinicopathological parameters in patients treated with breast conserving surgery based on univariate survival analysis. Stromal expression of legumain did not show any significant association with tumor recurrence.

Multivariate survival analysis showed that high expression of legumain in tumor cells was a poor prognostic factor for tumor recurrence in patients treated with breast conserving surgery independent of known other determinants of high risk DCIS including age at diagnosis, DCIS size, presentation, nuclear grade, comedo necrosis, margin status, molecular classes, and radiotherapy either for all recurrences ($HR=3.5$, 95% $CI=1.8-4.9$; $p=0.0003$) or when the analysis confined to invasive recurrences ($HR=3.4$, 95% $CI=1.8-8.3$; $p=0.002$) (Table 2 and Figure 4).

Interestingly, when legumain expression in tumor cells was incorporated with the other determinants of DCIS risk described by Van Nuys Prognostic Index (12), it provided better stratification for local recurrence risk, whereby high expression of legumain was associated with worse outcome in all risk groups when compared to similar groups with low legumain expression (Figure 5).

DISCUSSION

The underlying mechanisms promoting the transition from DCIS to invasive disease remain unclear and there is a demand to gain a better understanding. Several studies and risk assessment models are available; however, none is adequate for patients' risk stratification and hence a considerable percentage of patients with DCIS are either over- or under-treated. Furthermore, the biological and clinical heterogeneity of DCIS makes risk stratification quite challenging. An explanation of disease progression based exclusively on intrinsic tumor cell factors is insufficient, as there is a group of low grade DCIS with indolent appearance and low proliferation index that yet carries progression potential to invasive breast cancer (33). Studying the role of the DCIS microenvironment and the interaction between its various components and understanding how this influences disease behaviour could resolve the DCIS dilemma and provide a more adequate risk stratification model for personalised management (34-37). As invasion through the outer myoepithelial layer and basement membrane degradation is a key step in DCIS progression to invasive cancer, studying potential markers that drive this process and their prognostic value is a convincing approach to refine DCIS risk.

The lysosomal cysteine protease legumain is a proteolytic enzyme and plays role in autoimmunity and cancer (21, 38, 39). Overexpression of legumain is linked with poor prognosis in different tumors including invasive breast cancer (17, 22-24, 40, 41). Its action depends mainly on increasing the invasive and metastatic potential of the tumor via its proteolytic properties and stromal degradation (38). Comparing legumain expression between normal, borderline and invasive ovarian tissues reveals that it has a role not only in tumor migration and invasion but also in tumor development (40). However, similar studies are lacking in breast cancer to assess the role of legumain in DCIS. It was reported that legumain is differentially expressed between normal breast tissue and invasive breast cancer (16, 21, 39). Furthermore, using the METABRIC cohort for robust molecular data in a large number of invasive breast cancer, we have shown an association between aggressive behaviour of invasive breast cancer and higher levels of

legumain mRNA. These observations support our hypothesis that legumain is a promising candidate marker that requires additional studies to decipher its role in DCIS behaviour.

Here we explored the expression of legumain in a large well characterised cohort of DCIS and scored the protein expression in tumor cells and surrounding stromal fibroblasts. Interestingly, high legumain expression was associated with other features of high risk DCIS. These findings support the role of legumain in DCIS progression. Supporting this, our results showed that legumain expression is higher in DCIS co-existing with invasive carcinoma than pure DCIS, and much higher in the invasive component either within the tumor cells or in the surrounding stromal fibroblasts.

The poor prognostic value of legumain was shown with a shorter recurrence free interval in patients with high levels of legumain expression independently from other clinicopathological factors. These findings were consistent for all recurrent events, either DCIS or invasive breast cancer or when the analysis was confined to invasive recurrences only, which provides more evidence that legumain plays a key role in DCIS progression to invasive disease. Our study shows that expression of legumain in tumor epithelial cells, but not stromal cells, is associated with recurrence; a finding that might reflect the potential epithelial cell-intrinsic role of early stage tumors in extracellular matrix degradation that facilitates tumor progression and the dual role of tumor and stromal cells in progression and aggressiveness of advanced tumors. The latter interaction is supported by the dramatic increase of legumain expression in stromal cells surrounding the invasive component compared to those surrounding the DCIS component in mixed cases or those surrounding pure DCIS. However, further functional studies are highly recommended to understand the underlying mechanisms and functions of legumain expression in carcinogenesis and tumor progression either from the tumor cells or the surrounding stroma.

Incorporation of legumain with the other clinicopathological factors provided a better identification of different risk groups. These findings indicate that legumain is a

promising marker for better definition of high risk DCIS as well as for the identification of patients with lower risk where radiotherapy could be omitted.

Thus far, little is known about the biological processes which involve legumain in cancer progression. However, a correlation was observed between tumor invasion and metastasis and the presence of cysteine endopeptidases, such as cathepsins B and L (42). Protease zymogen cathepsins B and L may also be activated by legumain-mediated hydrolysis of asparaginyll bonds. Legumain acts as an asparaginyll endopeptidase in regulation of extracellular matrix remodelling through the activation of zymogen progelatinase A, which is an important mediator of extracellular matrix degradation, or the degradation of fibronectin, which is a main component of the extracellular matrix (43, 44). Animal tumor models generated with cells overexpressing legumain demonstrated an *in vivo* behaviour that is vigorous with more invasive growth and metastasis (39). This phenotype is proposed to result from the proteolytic function of legumain to activate other protease zymogens. The inhibitory effect of cystatins on tumor cells is consistent with the involvement of legumain, and perhaps other cysteine proteases, in tumor invasion and metastasis. Whether the tumor suppressing effect is mediated through inhibition of legumain catalytic activity or other cysteine proteases is presently unknown (43).

Legumain is present intracellularly in a pro-active form (38, 39) and one of the activating mechanisms is low pH. Interestingly, our findings showed that legumain is associated with presence of comedo type necrosis, which is consistent with low pH and supports our findings. Legumain is usually overexpressed in cells adjacent to necrosis (39), which was observed in our study as well where central cells facing the comedo necrosis showed higher legumain expression than the peripheral cells within the ducts.

The role of legumain in tumor aggressiveness is not related solely to its proteolytic activity but also to its proliferation activation mechanisms. This may be related to decreased apoptotic activity of cells and increased calcium influx into cells (21, 45). Supporting this possibility, our study showed that legumain was expressed in highly

proliferative DCIS, which may further augment the adverse action of legumain in the context of disease outcome.

The role of legumain in autoimmune disease and inflammatory process is undeniable (38). Legumain functions in antigen presentation to inflammatory cells may be a cause for such phenomenon. Overexpression of legumain in tumor associated macrophages and endothelial cells of the surrounding tissues has been reported (41). Accordingly, the link between legumain and dense inflammatory cells infiltrates is warranted to be investigated. We previously reported that dense tumor infiltrating lymphocytes have poor prognostic significance in DCIS, a reverse phenomenon to the invasive disease for which the underlying mechanisms are unclear (26). We saw a striking association of high stromal legumain and a dense lymphocytic infiltrate in pure DCIS (Table 1) that may be associated with an inflammatory function for legumain. Taken together, legumain may interact with the inflammatory cascade and affect DCIS behaviour.

Conclusion

Extracellular matrix degradation is an essential step for DCIS progression to invasive disease. Legumain might have a potential role in DCIS aggressiveness through its proteolytic activity and regulatory mechanism in cellular proliferation. Additional functional studies to decipher the role of legumain and its mechanism of action in DCIS behaviour are warranted. Legumain may also be a valuable prognostic indicator especially for invasive recurrence.

Limitations of the study

This study has been carried out on TMA sections, which might underestimate the role of tumour heterogeneity. However, all cases in our cohort were histologically reviewed before TMA construction and used multiple cores for cases with heterogeneous grades or morphological patterns. Moreover, our cohort did not include any patients treated with endocrine therapy.

Acknowledgment:

This research was supported and funded by the Egyptian Ministry of Higher Education and Scientific Research. We thank the Nottingham Health Science Biobank and Breast Cancer Now Tissue Bank for the provision of tissue samples

References

1. Pang JM, Gorringer KL, Fox SB. Ductal carcinoma in situ - update on risk assessment and management. *Histopathology* 2016;68:96-109.
2. Bartlett JM, Nofech-Moses S, Rakovitch E. Ductal carcinoma in situ of the breast: can biomarkers improve current management? *Clin Chem*. 2014;60:60-7.
3. Sanger N, Engels K, Graf A, *et al*. Molecular Markers as Prognostic Factors in DCIS and Small Invasive Breast Cancers. *Geburtshilfe und Frauenheilkunde* 2014;74:1016-22.
4. Knopfmacher A, Fox J, Lo Y, Shapiro N, Fineberg S. Correlation of histopathologic features of ductal carcinoma in situ of the breast with the oncotype DX DCIS score. *Mod Pathol*. 2015;28:1167-73.
5. Rakovitch E, Nofech-Moses S, Hanna W, *et al*. A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. *Breast Cancer Res Treat*. 2015;152:389-98.
6. Solin LJ, Gray R, Baehner FL, *et al*. A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. *J Natl Cancer Inst*. 2013;105:701-10.
7. Yi M, Meric-Bernstam F, Kuerer HM, *et al*. Evaluation of a breast cancer nomogram for predicting risk of ipsilateral breast tumor recurrences in patients with ductal carcinoma in situ after local excision. *J Clin Oncol*. 2012;30:600-7.
8. Rudloff U, Jacks LM, Goldberg JL, *et al*. Nomogram for predicting the risk of local recurrence after breast-conserving surgery for ductal carcinoma in situ. *J Clin Oncol*. 2010;28:3762-9.
9. Mazouni C, Delaloge S, Rimareix F, Garbay JR. Nomogram for risk of relapse after breast-conserving surgery in ductal carcinoma in situ. *J Clin Oncol*. 2011;29:e44; author reply e5-6.
10. Lee SK, Yang JH, Woo SY, Lee JE, Nam SJ. Nomogram for predicting invasion in patients with a preoperative diagnosis of ductal carcinoma in situ of the breast. *Br J Surg*. 2013;100:1756-63.
11. MacAusland SG, Hepel JT, Chong FK, *et al*. An attempt to independently verify the utility of the Van Nuys Prognostic Index for ductal carcinoma in situ. *Cancer* 2007;110:2648-53.
12. Boland GP, Chan KC, Knox WF, Roberts SA, Bundred NJ. Value of the Van Nuys Prognostic Index in prediction of recurrence of ductal carcinoma in situ after breast-conserving surgery. *Br J Surg*. 2003;90:426-32.
13. Yeong J, Thike AA, Tan PH, Iqbal J. Identifying progression predictors of breast ductal carcinoma in situ. *J Clin Pathol*. 2017;70:102-8.
14. Gonzalez LO, Gonzalez-Reyes S, Junquera S, *et al*. Expression of metalloproteases and their inhibitors by tumor and stromal cells in ductal carcinoma in situ of the breast and their relationship with microinvasive events. *J Cancer Res Clin Oncol*. 2010;136:1313-21.

15. Coussens LM, Fingleton B, Matrisian LM. Matrix metalloproteinase inhibitors and cancer: trials and tribulations. *Science* 2002;295:2387-92.
16. Gawenda J, Traub F, Luck HJ, Kreipe H, von Wasielewski R. legumain expression as a prognostic factor in breast cancer patients. *Breast Cancer Res Treat.* 2007;102:1-6.
17. Guo P, Zhu Z, Sun Z, *et al.* Expression of legumain correlates with prognosis and metastasis in gastric carcinoma. *PloS One* 2013;8:e73090.
18. Chen J, Fortunato M, Stevens R, Barrett A. Activation of progelatinase A by mammalian legumain, a recently discovered cysteine proteinase. *Biol Chem.* 2001;382:777-83.
19. Mai J, Finley RL, Waisman DM, Sloane BF. Human procathepsin B interacts with the annexin II tetramer on the surface of tumor cells. *J Biol Chem.* 2000;275:12806.
20. Stern L, Perry R, Ofek P, *et al.* A novel antitumor prodrug platform designed to be cleaved by the endoprotease legumain. *Bioconj Chem.* 2009;20:500.
21. Zhen Y, Chunlei G, Wenzhi S, *et al.* Clinicopathologic significance of legumain overexpression in cancer: a systematic review and meta-analysis. *Sci Rep.* 2015;5:16599.
22. Wu M, Shao GR, Zhang FX, *et al.* legumain protein as a potential predictive biomarker for Asian patients with breast carcinoma. *Asian Pac J Cancer Prev.* 2014;15:10773-7.
23. Murthy RV, Arbman G, Gao J, Roodman GD, Sun XF. legumain expression in relation to clinicopathologic and biological variables in colorectal cancer. *Clin Cancer Res.* 2005;11:2293-9.
24. Ohno Y, Nakashima J, Izumi M, *et al.* Association of legumain expression pattern with prostate cancer invasiveness and aggressiveness. *World J Urol.* 2013;31:359-64.
25. Miligy IM, Gorringe KL, Toss MS, *et al.* Thioredoxin-interacting protein is an independent risk stratifier for breast ductal carcinoma in situ. *Mod Pathol.* 2018.
26. Toss MS, Miligy I, Al-Kawaz A, *et al.* Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma in situ of the breast. *Mod Pathol.* 2018.
27. Gilkes DM, Chaturvedi P, Bajpai S, *et al.* Collagen prolyl hydroxylases are essential for breast cancer metastasis. *Cancer Res.* 2013;73:3285-96.
28. Pan PW, Zhang Q, Bai F, Hou J, Bai G. Profiling and comparative analysis of glycoproteins in Hs578BST and Hs578T and investigation of prolyl 4-hydroxylase alpha polypeptide II expression and influence in breast cancer cells. *Biochemistry (Mosc)* 2012;77:539-45.
29. Xiong G, Deng L, Zhu J, Rychahou PG, Xu R. Prolyl-4-hydroxylase α subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition. *BMC cancer* 2014;14:1.
30. Hendry S, Pang J-MB, Byrne D, *et al.* Relationship of the breast ductal carcinoma in situ immune microenvironment with clinico-pathological and genetic features. *Clin Cancer Res.* 2017.

31. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res.* 2004;10:7252-9.
32. Curtis C, Shah SP, Chin SF, *et al.* The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature* 2012;486:346-52.
33. Ma XJ, Dahiya S, Richardson E, Erlander M, Sgroi DC. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.* 2009;11:R7.
34. Allen MD, Thomas GJ, Clark S, *et al.* Altered microenvironment promotes progression of preinvasive breast cancer: myoepithelial expression of α v β 6 integrin in DCIS identifies high-risk patients and predicts recurrence. *Clin Cancer Res.* 2014;20:344-57.
35. Semenza GL. The hypoxic tumor microenvironment: A driving force for breast cancer progression. *Biochim Biophys Acta* 2016;1863:382-91.
36. Thompson E, Taube JM, Elwood H, *et al.* The immune microenvironment of breast ductal carcinoma in situ. *Mod Pathol* 2016;29:249-58.
37. Allinen M, Beroukhi R, Cai L, *et al.* Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 2004;6:17-32.
38. Dall E, Brandstetter H. Activation of legumain involves proteolytic and conformational events, resulting in a context- and substrate-dependent activity profile. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2012;68:24-31.
39. Liu C, Sun C, Huang H, Janda K, Edgington T. Overexpression of legumain in tumors is significant for invasion/metastasis and a candidate enzymatic target for prodrug therapy. *Cancer research* 2003;63:2957-64.
40. Wang L, Chen S, Zhang M, *et al.* Legumain: a biomarker for diagnosis and prognosis of human ovarian cancer. *J Cell Biochem.* 2012;113:2679-86.
41. Luo Y, Zhou H, Krueger J, *et al.* Targeting tumor-associated macrophages as a novel strategy against breast cancer. *Journal Clin Invest.* 2006;116:2132-41.
42. Duivenvoorden HM, Rautela J, Edgington-Mitchell LE, *et al.* Myoepithelial cell-specific expression of stefin A as a suppressor of early breast cancer invasion. *J Pathol* 2017;243:496-509.
43. Cheng T, Hitomi K, van Vlijmen-Willems IM, *et al.* Cystatin M/E is a high affinity inhibitor of cathepsin V and cathepsin L by a reactive site that is distinct from the legumain-binding site. A novel clue for the role of cystatin M/E in epidermal cornification. *J Biol Chem.* 2006;281:15893-9.
44. Chen JM, Fortunato M, Stevens RA, Barrett AJ. Activation of progelatinase A by mammalian legumain, a recently discovered cysteine proteinase. *Biol Chem.* 2001;382:777-83.
45. Andrade V, Guerra M, Jardim C, *et al.* Nucleoplasmic calcium regulates cell proliferation through legumain. *J Hepatol* 2011;55:626-35.

Table 1: Correlation between legumain expression with different clinicopathological parameters in the pure DCIS cohort.

Clinicopathological Parameters	Legumain expression in tumor epithelial cells		χ^2 (p-value)	Legumain expression in stromal fibroblasts		χ^2 (p-value)
	Low (N=359) N. (%)	High (N=105) N. (%)		Low (N=259) N. (%)	High (N=205) N. (%)	
Age (years) ≤50 >50	88 (25) 271 (75)	27 (24) 78 (76)	0.1 (0.802)	62 (24) 197 (76)	53 (26) 152 (74)	0.3 (0.635)
Presentation Screening Symptomatic	177 (49) 182 (51)	48 (46) 57 (54)	0.4 (0.517)	134 (51) 125 (49)	91 (44) 114 (56)	2.5 (0.116)
DCIS Size (mm) <16 16-40 >40	112 (31) 148 (41) 98 (28)	22 (21) 45 (43) 37 (36)	4.8 (0.092)	77 (30) 107 (41) 75 (29)	57 (28) 86 (42) 60 (30)	0.2 (0.927)
Nuclear Grade Low Moderate High	51 (14) 113 (32) 195 (54)	8 (8) 15 (14) 82 (78)	19.2 (<0.0001)	39 (15) 84 (32) 136 (53)	20 (10) 44 (21) 141 (69)	12.6 (0.002)
Comedo necrosis Yes No	226 (63) 133 (37)	80 (76) 25 (23)	6.3 (0.012)	160 (62) 99 (38)	59 (30) 146 (70)	4.5 (0.033)
Estrogen receptor Negative Positive	62 (19) 266 (81)	54 (54) 47 (46)	46.8 (<0.0001)	42 (18) 195 (82)	74 (39) 118 (62)	23.3 (<0.0001)
Progesterone receptor Negative Positive	114 (34) 217 (66)	67 (66) 35 (34)	31.3 (<0.0001)	76 (32) 165 (68)	105 (55) 87 (45)	23.5 (<0.0001)
HER2 status Negative Positive	256 (79) 68 (21)	60 (63) 35 (37)	9.9 (0.002)	190 (80) 47 (20)	126 (69) 56 (31)	6.6 (0.010)
Proliferation index (Ki-67) High Low	59 (20) 240 (80)	36 (37) 62 (63)	11.7 (0.001)	31 (33) 184 (67)	64 (61) 118 (40)	23.3 (<0.0001)
Molecular classes Luminal A Luminal B HER2 Enriched Triple negative	160 (58) 61 (22) 25 (9) 29 (11)	22 (24) 18 (20) 22 (25) 28 (31)	46.4 (<0.0001)	127 (63) 35 (18) 19 (10) 18 (9)	55 (33) 44 (26) 28 (17) 39 (24)	36.3 (<0.0001)
Tumour infiltrating lymphocytes Dense Sparse	123 (44) 159 (56)	55 (63) 33 (37)	9.6 (0.002)	72 (36) 130 (64)	106 (63) 62 (37)	27.7 (<0.0001)
DCIS Type* Pure DCIS DCIS with invasive breast cancer	359 (75) 123 (25)	105 (61) 68 (39)	11.7 (0.001)	259 (91) 27 (9)	205 (56) 164 (44)	95.6 (<0.0001)

Significant p values are in bold

DCIS; ductal carcinoma *in situ*, HER2; Human epidermal growth factor receptor 2, n; number

*Including the cases in both cohorts; i.e. pure DCIS cohort (n=464) + DCIS-mixed cohort (n=191).

Table 2: Multivariate survival analysis (Cox regression model) of variables predicting outcome in terms of ipsilateral local recurrence in patients treated by breast conserving surgery in pure DCIS cohort

A) All recurrences

Parameters	Hazard ratio (HR)	95% confidence interval (CI)		p-value
		Lower	Upper	
High legumain expression	3.5	1.8	4.9	0.0003
Patient Age	1.4	1.2	2.6	0.019
DCIS presentation	1.8	0.8	3.8	0.103
DCIS size	0.8	0.5	1.3	0.297
DCIS nuclear Grade	1.5	0.9	2.6	0.148
Comedo necrosis	0.7	0.3	1.4	0.268
Molecular classes	0.8	0.6	1.2	0.260
Radiotherapy	0.3	0.1	0.9	0.045
Margin status	0.8	0.4	1.3	0.752

B) Invasive recurrence

Parameters	Hazard ratio (HR)	95% confidence interval (CI)		p-value
		Lower	Upper	
High legumain expression	3.4	1.8	8.3	0.002
Patient Age	1.5	0.7	2.3	0.129
DCIS presentation	1.4	0.5	3.6	0.511
DCIS size	1.2	0.6	2.4	0.688
DCIS nuclear Grade	1.4	0.7	2.9	0.349
Comedo necrosis	1.1	0.4	2.8	0.912
Molecular classes	0.5	0.3	0.9	0.027
Radiotherapy	0.2	0.1	0.9	0.040
Margin status	0.9	0.4	1.9	0.813

Significant *p* values are in bold

Supplementary Table 1: Characterization of pure DCIS cohort (n=464)

Clinicopathological Parameters	No. of cases (%)
Age (years) ≤50 >50	115 (25) 349 (75)
Presentation Screening Symptomatic	225 (49) 239 (51)
DCIS Size (mm) <16 16-40 >40	134 (29) 193 (42) 135 (29)
Nuclear Grade Low Moderate High	59 (12) 128 (28) 277 (60)
Comedo necrosis* Yes No	306 (66) 158 (34)
Estrogen receptor Negative Positive	116 (27) 313 (73)
Progesterone receptor Negative Positive	181 (42) 252 (58)
HER2 status Negative Positive	316 (75) 103 (25)
Proliferation index (Ki-67) High (>14%) Low (≤14%)	95 (24) 302 (76)
Molecular classes Luminal A Luminal B HER2 Enriched Triple negative	182 (50) 79 (22) 47 (13) 57 (15)
Tumour infiltrating lymphocytes Dense Sparse	178 (48) 192 (52)
Treatment Mastectomy Breast conserving surgery Breast conserving surgery+ adjuvant Radiotherapy	259 (56) 137 (30) 68 (14)
Margin status** Positive (tumour on ink) <2mm free margin ≥2mm free margin	5 (2) 6 (2) 194 (96)

*Comedo necrosis defined definite central necrosis within the DCIS ducts that have been identified as necrosis with nuclear debris (we did not use % as this was subjective). Cases showing apoptotic bodies or single cell necrosis were not considered as comedo necrosis.

**Based on new guidelines by the American Society of Clinical Oncology and the Society of Surgical Oncology 2016. Data shown here is for cases treated with BCS (n=205)

Supplementary Table 2: Correlation between legumain expression and different clinicopathological parameters in the pure DCIS cohort using continuous data.

Clinicopathological Parameters	Number of cases	Legumain expression in tumor epithelial cells		Legumain expression in stromal fibroblasts	
		Mean Rank	<i>p</i> -value	Mean Rank	<i>p</i> -value
Age (years) ≤50 >50	115 349	225.1 234.9	0.488	243.4 228.9	0.276
Presentation Screening Symptomatic	225 239	228.8 235.9	0.563	221.7 242.6	0.070
DCIS Size (mm) <16 16-40 >40	134 193 135	211.7 231.6 250.9	0.054	224.9 231.8 237.5	0.701
Nuclear Grade Low Moderate High	59 128 277	198.9 201.1 254.1	<0.0001	207.1 208.5 248.9	0.002
Comedo necrosis Yes No	306 158	247.4 203.5	0.001	242.5 213.1	0.016
Estrogen receptor Negative Positive	116 313	277.5 191.9	<0.0001	262.9 197.3	<0.0001
Progesterone Receptor Negative Positive	181 252	259.9 186.2	<0.0001	250.7 192.8	<0.0001
HER2 status Negative Positive	316 103	195.9 253.1	<0.0001	200.9 238.0	0.003
Proliferation index (Ki-67) High Low	95 302	245.3 184.4	<0.0001	244.2 184.7	<0.0001
Molecular classes Luminal A Luminal B HER2 Enriched Triple negative	182 79 47 57	146.3 195.4 237.4 238.2	<0.0001	151.9 201.6 214.3 230.6	<0.0001
Tumour infiltrating lymphocytes Dense Sparse	178 192	207.5 165.1	<0.0001	216.3 156.9	<0.0001
DCIS Type Pure DCIS DCIS with invasive breast cancer	464 191	319.2 349.4	0.049	283.2 436.8	<0.0001

Significant *p* values are in bold

DCIS; ductal carcinoma *in situ*, HER2; Human epidermal growth factor receptor 2

Supplementary Table 3: Correlation between *legumain* mRNA level and the clinicopathological parameters in the METABRIC series of invasive breast cancers (n=1980).

Clinicopathological parameters	Number of cases	Mean <i>Legumain</i> mRNA level	<i>p</i> -value
Patient Age (years)			
<50	383	8.7	0.746
≥50	1556	8.6	
Tumour Size (mm)			
≤20	622	8.7	0.443
>20	1331	8.5	
Histologic Grade			
1	170	8.2	0.034
2	770	8.2	
3	952	8.3	
Lymph node metastasis			
Negative	1035	8.5	0.040
Positive	938	8.6	
Estrogen Receptor Status			
Positive	1506	8.4	0.001
Negative	474	8.9	
HER2 Status			
Negative	1733	8.5	0.006
Positive	247	8.8	
PAM50 molecular classes			
Luminal A	718	8.4	<0.0001
Luminal B	488	8.5	
Basal-like	329	9.0	
HER2 enriched	240	8.7	
Normal like	199	8.9	

Significant *p* values are in bold

METABRIC; Molecular Taxonomy of Breast Cancer International Consortium

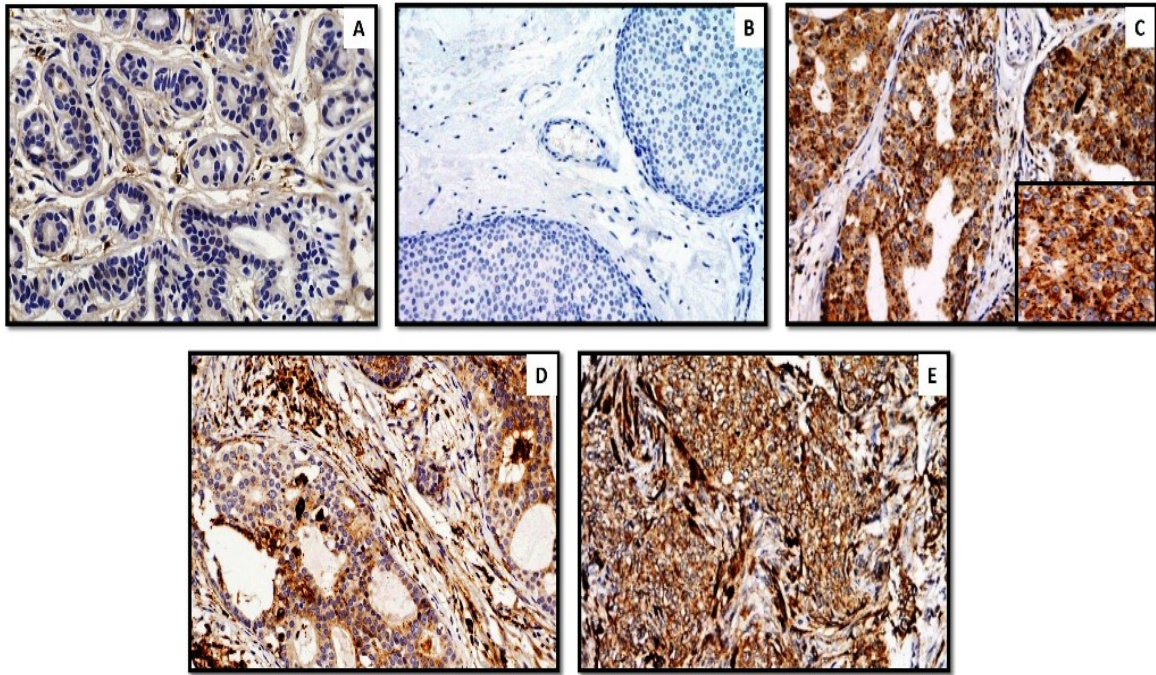


Figure 1: A) Normal breast ductolobular unit (x20) shows negative staining of legumain; B) Negative legumain expression (x20) in a pure DCIS case; C) strong expression of legumain in tumor cells and surrounding fibroblasts (x20) in a pure DCIS case (Inset: high power view showing the granular pattern of legumain expression). D) High expression of legumain in the fibroblasts surrounding DCIS case (x40). E) Expression of legumain in a mixed case (x40) showing strong staining in invasive component either within the tumor cells or surrounding stromal fibroblasts.

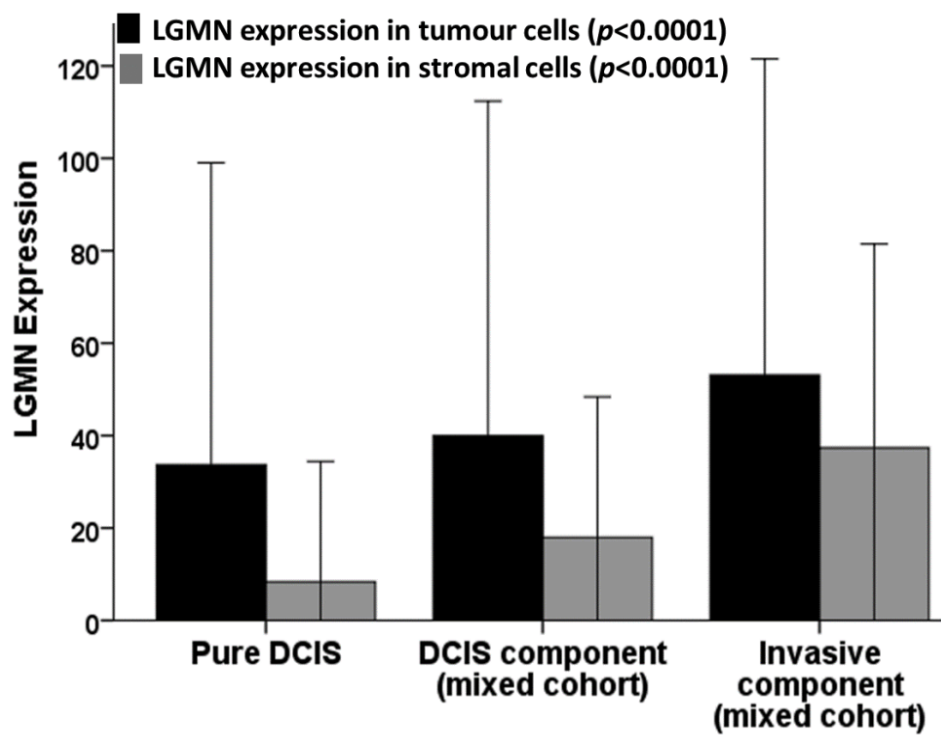


Figure 2: Bar chart showing differences of legumain expression between pure DCIS and DCIS-mixed both in tumor cells (black bars) and surrounding stroma (grey bars). P-value from ANOVA, Error bars represent +2 standard deviation.

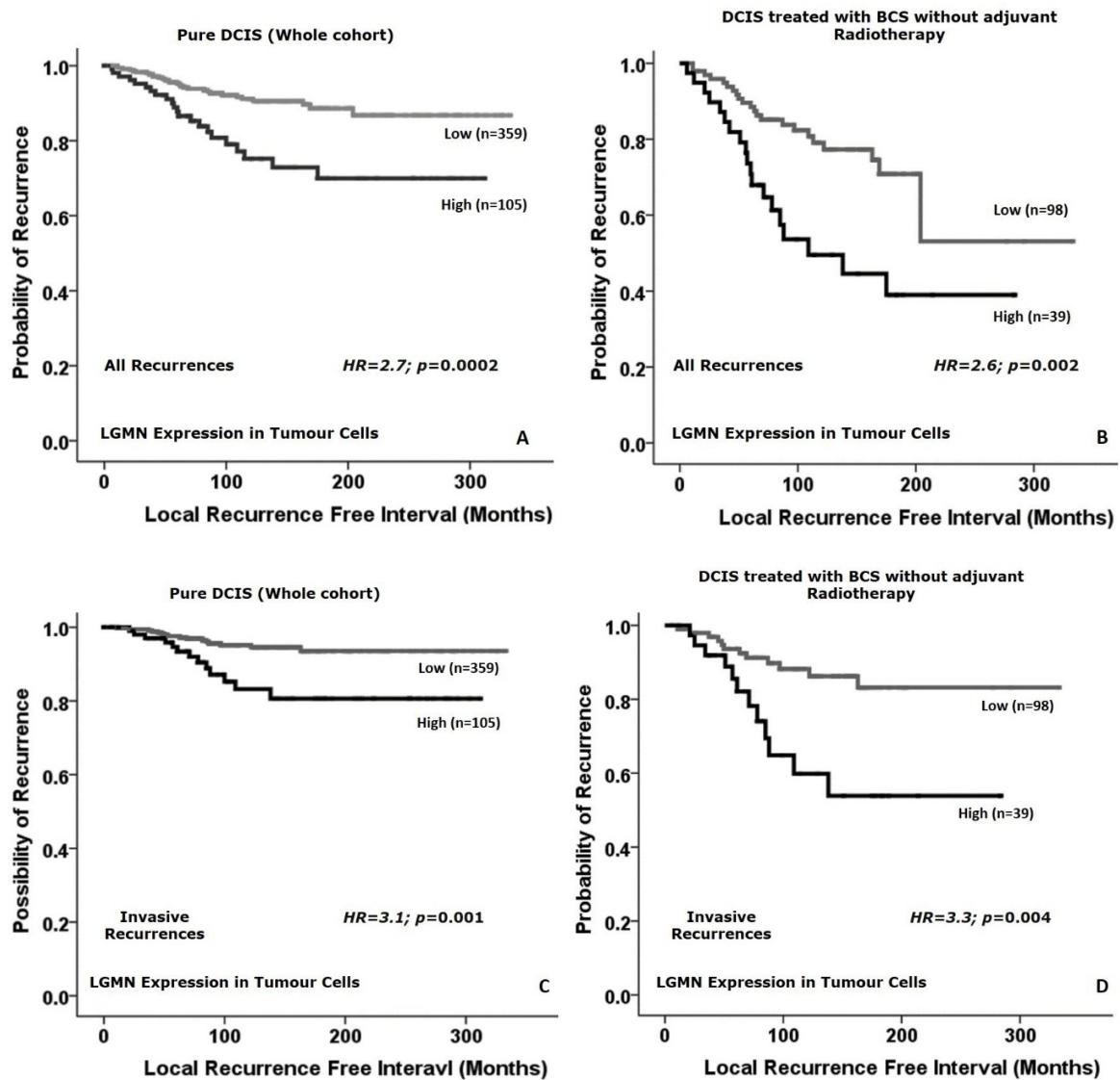


Figure 3: Kaplan Meier curves show that high expression of legumain within the tumor epithelial cells is associated with shorter ipsilateral local recurrence free interval in the whole series (A), and in breast conserving surgery (BCS) without adjuvant radiotherapy (B). High expression also showed an association with shorter local recurrence free interval as invasive disease in the whole series (C) and in patients treated with breast conserving surgery without adjuvant radiotherapy (D).

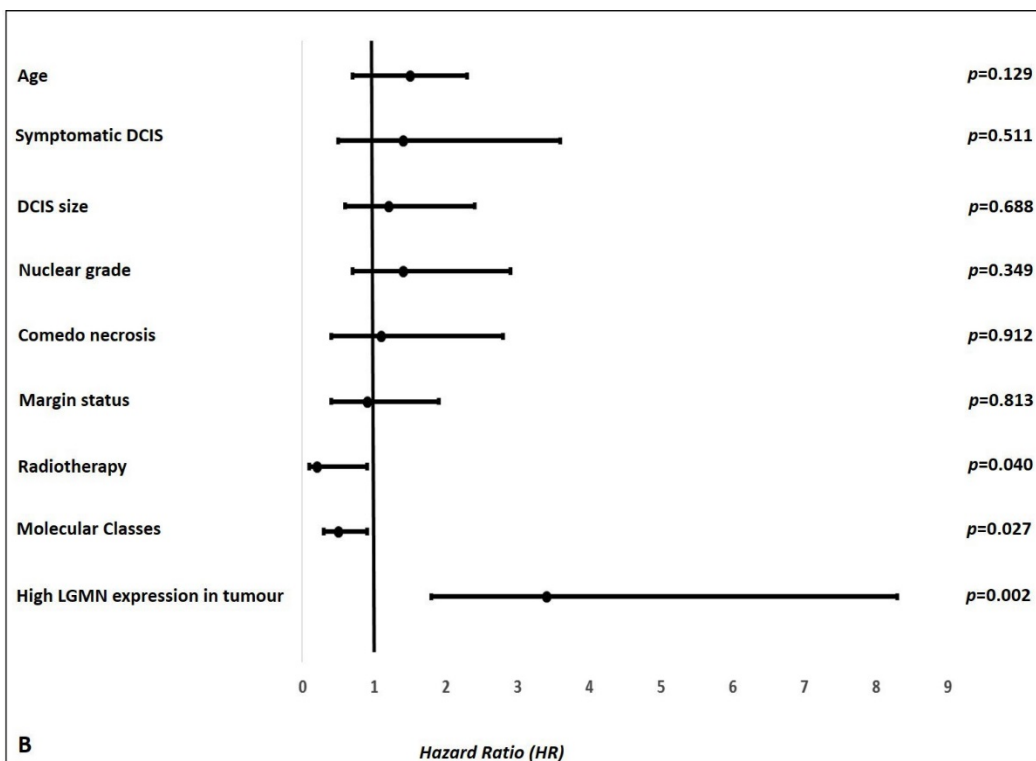
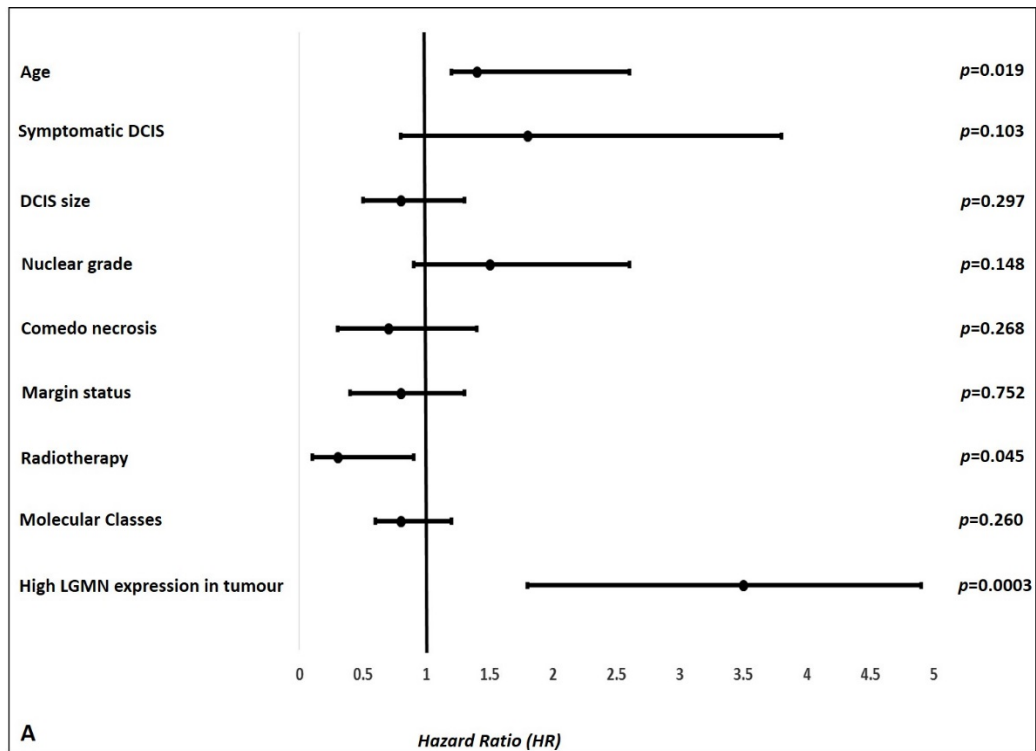


Figure 4: Forest plots showing the hazard ratio of the different clinicopathological parameters and ipsilateral tumor recurrence for patients treated with breast conserving surgery in pure DCIS cohort based on the multivariate analysis results for; A) all recurrences whether DCIS or invasive breast cancer and B) for invasive recurrences only.

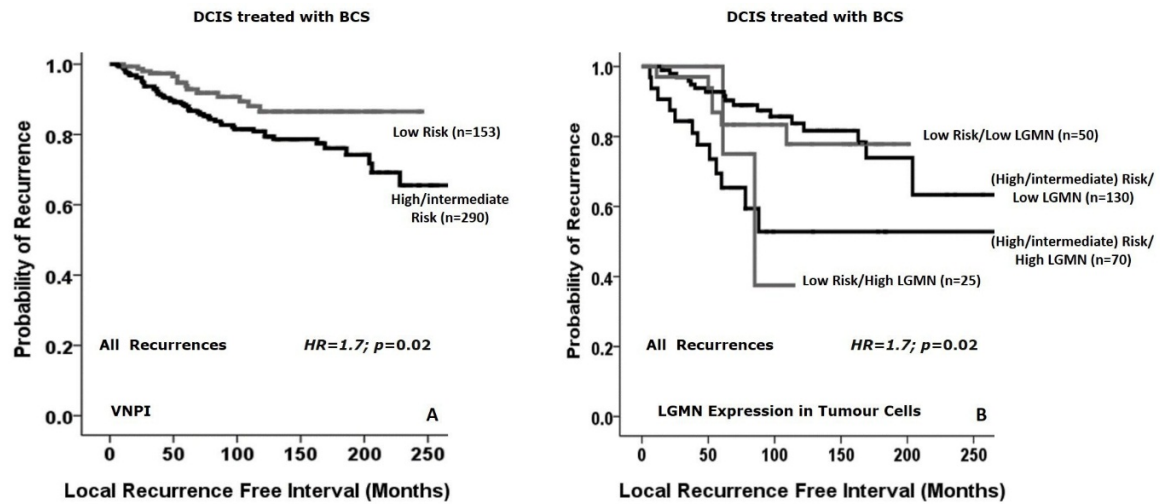
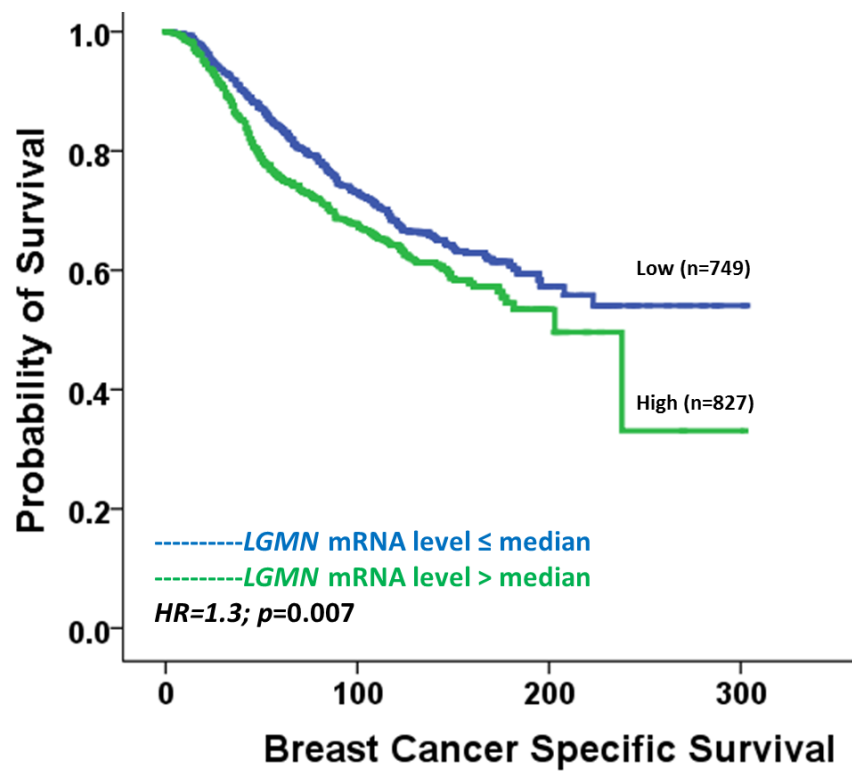
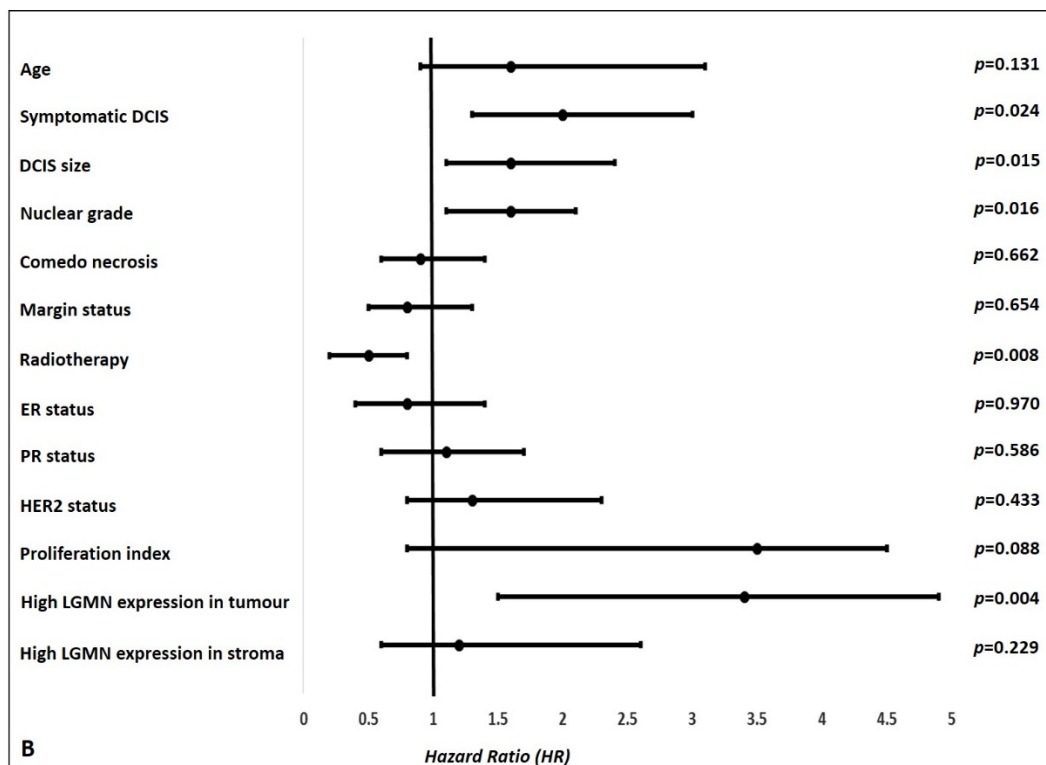
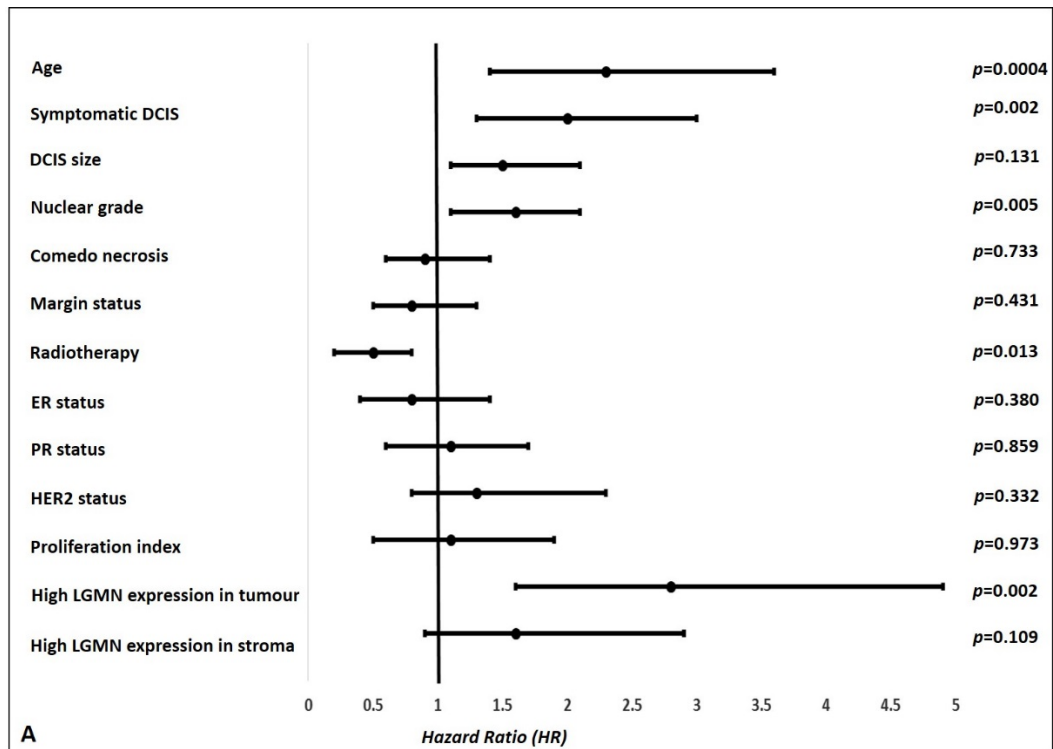


Figure 5: Kaplan Meier curves show the association between DCIS risk and local recurrence free interval in patients treated with breast conserving surgery based on Van Nuys Prognostic Index alone (A), and when legumain was incorporated with the Van Nuys Prognostic Index (B).



Supplementary Figure 1: Association between *legumain* mRNA level and outcome in terms of breast cancer specific survival (BCSS) in the METABRIC series. The cohort was split into high and low mRNA expression based on the median (8.59).



Supplementary Figure 2: Forest plots showing the univariate analysis results of association between different clinicopathological parameters and ipsilateral tumor recurrence for patients treated with breast conserving surgery in pure DCIS cohort; A) all recurrences whether DCIS or invasive and B) for invasive recurrences only. High expression of legumain in tumor epithelial cells is associated with higher recurrence risk in both groups.